Article:
Kharroubi, SA, Edlin, R, Meads, D orcid.org/0000-0003-1369-2483 et al. (1 more author) (2018) Bayesian statistical models to estimate EQ-5D utility scores from EORTC QLQ data in myeloma. Pharmaceutical Statistics, 17 (4). pp. 358-371. ISSN 1539-1604

https://doi.org/10.1002/pst.1853

Title: Bayesian statistical models to estimate EQ-5D utility scores from EORTC QLQ data in Myeloma

Samer A Kharroubi[1] PhD, Richard Edlin[2] PhD, David Meads[3] PhD, Christopher McCabe[4] PhD

1. Department of Nutrition and Food Sciences, Faculty of Agricultural and Food Sciences, American University of Beirut, Lebanon
2. School of Population Health, University of Auckland, New Zealand
3. Academic Unit of Health Economics, University of Leeds, Leeds (UK)
4. Department of Emergency Medicine, Faculty of Medicine and Dentistry, University of Alberta, Edmonton (Canada)

Corresponding Author: Samer A. Kharroubi, Department of Nutrition and Food Sciences, Faculty of Agricultural and Food Sciences, American University of Beirut, Lebanon

Tel: +961 (0) 1 350 000 Ext 4541

Fax: +961 (0) 1 744 460

Email: sk157@aub.edu.lb

Source of financial support: The work was supported by the Medical Research Council [MRC MRP 06/92/63].

Keywords: Bayesian methods; EQ-5D; Multiple Myeloma; Quality of Life; two-part models; Cost-utility analysis; MCMC.

Running head: Bayesian estimation for QoL data

Paper for submission to Pharmaceutical Statistics
Wednesday, 02 May 2018

Title: Bayesian statistical models to estimate EQ-5D utility scores from EORTC QLQ data in Myeloma

Abstract

Background: It is well documented that the modelling of health related quality of life (HRQoL) data is difficult as the distribution of such data is often strongly right/left skewed and it includes a significant percentage of observations at one. **Objectives:** To develop a series of two-part models (TPM) that deal with these issues. **Methods:** Data from the UK Medical Research Council (MRC) Myeloma IX trial were used to examine the relationship between the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30/QLQ-MY20 scores and the European QoL-5 Dimensions (EQ-5D) utility score. Four different TPM models were developed. The models fitted included TPM with normal regression, TPM with normal regression with variance a function of participant characteristics, TPM with log-transformed data and TPM with gamma regression and a log link. The cohort of 1839 patients was divided into 75% derivation sample, to fit the different models, and 25% validation sample to assess the predictive ability of these models by comparing predicted and observed mean EQ-5D scores in the validation set, unadjusted R^2 and Root Mean square Error (RMSE). **Results:** Predictive performance in the derivation dataset depended on the criterion used, with R^2/adjusted-R^2 favouring the TPM model with normal regression and mean predicted error favouring the TPM model with gamma regression. The TPM model with gamma regression performs best within the validation dataset under all criteria. **Conclusions:** TPM regression models provide flexible approaches to estimate mean EQ-5D utility weights from the EORTC QLQ-C30/QLQ-MY20 for use in economic evaluation.

Key words: Bayesian methods; EQ-5D; Multiple Myeloma; Quality of Life; two-part models; Cost-utility analysis; MCMC.

1. Introduction

Preference-based measures of health related quality of life (HRQoL) are widely used to calculate quality adjusted life years (QALYs) for use in cost effectiveness analyses. Examples of these are the European QoL-5 Dimensions (EQ-5D) questionnaire [1], the six-dimensional health state short form (SF-6D) [2] and Health Utility Index 2 (HUI-II) [3] instruments. QALYs are formed as a weighted sum of life expectancy, where each year is weighted on a HRQoL scale that reflect the sacrifices that people would make to attain or avoid the particular states in those years. In such measures, zero and one are assigned specific meanings. At a HRQoL value of one, a person would not make any sacrifices for better health and they are said to be in 'full health'. In contrast, those with a HRQoL value of zero occupy a state 'as bad as dead', in the sense that they would swap a 100% certainty of death for remaining in that state; once dead, a HRQoL value of zero is also assigned.

As a HRQoL value of one represents the state where a person no longer sacrifices to obtain better health, HRQoL does not usually exceed one and often has a noticeable 'spike' here. As most people are only willing to make moderate sacrifices for health, HRQoL values lie tend to at the higher end of the measurement scale, with some observations displaying extremely low levels of HRQoL. In principle, there is no lower bound to a HRQoL value, so that support may include large negative values. Thus, HRQoL values often have a noticeable density spike at one and a negatively-skewed distribution. In contrast the disutility of a health state (formed as one minus the HRQoL value), has a significant percentage of zero-disutility values and is positively skewed. The disutility of a particular health state therefore shares the same characteristics as cost data, including heteroskedasticity.

Most studies modelling HRQoL data use linear regression assuming normal and homooscedastic error terms [4-6]. The latter condition is unlikely to be satisfied for bounded variables for mean values close to bounds [7]. Alternative regression methods that model HRQoL data include censored least absolute deviation (CLAD) models [8, 9], Tobit models [8, 10] and median regression [9]. Of these, CLAD and Tobit methods model an underlying latent variable, which is censored at one. Such models are not

necessarily appropriate for HRQoL data because preferences are measured on a scale where values cannot be over 1 ('no HRQoL deficit') [8]. Median regression also addresses the non-normality of the data [11]. However, the regression usually concentrates on the mean rather than the median when HRQoL data are regressed on individual covariates to predict QALYs [8-10].

One approach advocated to address the problem generated by zero/one HRQoL observations is to apply a two-part model (TPM) [12-15] on disutility. Such an approach fits two separate models. Firstly, a logistic regression model is used to predict whether patients indicate any HRQoL deficit. Secondly, for those individuals who indicate nonzero disutility, a regression model is fitted to estimate the magnitude of the deficit. This approach is broadly similar to existing approaches dealing with cost data. For example, Cooper et al [13] have proposed a Bayesian TPM in which a logistic regression model was used first to predict the conditional probability of observing zero costs in the sample, followed by a linear regression model fitted to the log-transformed cost applied to those individuals reporting positive costs. This paper, and similar approaches elsewhere [16-17] provide flexible approaches to regress the mean of an outcome with truncated support such as HRQoL on covariates

This paper focuses on the application of four different TPM specifications of increasing complexity to predict HRQoL utility in myeloma patients. The paper is organised with the next section describing the motivating data for the models described herein. Then Section 3 outlines the models used for the analysis including assessment of model complexity, model fit and assessment of model prediction. Section 4 illustrates the application of the proposed method to the analysis of HRQoL data from the case study presented in Section 2. Finally, the implications of the results and directions for future research conclude the paper.

2. Motivating data set

2.1 Measures

The EuroQol 5D is recommended for use in economic evaluations by National Institute for Health and Care Excellence (NICE) [18], with the EQ-5D-3L [19] used here. This measure classifies patients into one of 243 health states varying in five dimensions

(alongside additional states for ‘dead’ and ‘unconscious’): mobility; self-care; usual activities; pain/ discomfort; and anxiety/depression. The levels of each dimension are roughly: none, moderate, and severe. The EQ-5D is of demonstrated validity and reliability [1, 20] and population values are available for both the UK [21] and elsewhere [22].

The European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ-C30) is a commonly-used instrument for measuring general cancer quality of life. It has been translated into more than 65 languages and is used widely internationally [23]. It covers several health domains, as well as cancer-specific symptoms of disease, the side effects of treatment, psychological distress, physical functioning, social interaction, global health, and quality of life. Most of the questions have a categorical response (Not at all; A little; Quite a bit; and Very much), with two questions relying on the use of a Visual Analogue Scale (VAS). The raw questionnaire responses are transformed to produce scores (0-100) on a set of five function scales (physical, role, emotional, cognitive, and social functioning) and nine symptom scales, along with a scale representing global quality of life. Higher scores indicate better functioning and more severe symptoms on the functioning and symptom scales, respectively. The EORTC QLQ-C30 scale has undergone extensive psychometric testing [24].

The myeloma cancer module QLQ-MY20 is acceptable for use among patients varying in disease level and treatment modality (i.e. surgery, chemotherapy, radiotherapy and hormonal treatment). It should always be used as a complement to the QLQ-C30. The myeloma module is designed to assess the symptoms and side effects of treatment and their impact on everyday life [25]. The module comprises 20 questions addressing four domains of QoL important in myeloma: body image, diseases symptoms, treatment side effects and future perspective. The module was developed according to the guidelines, and approved after formal review. As with the EORTC QLQ-C30 questions, the QLQ-MY20 questions have a hierarchical response with ordinal scores transformed to a 0-100 scale and each domain analysed separately. Both QLQ-C30 and –MY20 domain scores are re-transformed to 0-1 scale, and for this to allow values to mirror the overall potential contribution to quality of life.

2.2. Data source

The EORTC QLQ-C30 and QLQ-MY20 instruments are mapped onto the EQ-5D instrument using data from MYELOMA-IX[1] [26]. This randomized controlled trial of the effectiveness of new treatment regimens at both diagnosis (CVAD vs. C-Thal-Dex or C-Thal-Dex attenuated; sodium clodrinate vs zoledronic acid) and maintenance (no therapy vs thalidomide). The trial needed a sample size of 1600 patients enlisted across hospitals in the UK, New Zealand and South Africa over 5 years. Overall, the trial recruited 1839 patients, who provided a total of 3184 observations.

3. Model development and validation

In this section, we develop a series of TPMs that take into account the typical characteristics of HRQoL data. All models shown are implemented from a Bayesian perspective using Gibbs sampling MCMC methods freely available in the specialist software WinBUGS [27]. The WinBUGS code is available from the corresponding author. MCMC methods permit great flexibility in the specification of complex non-standard models and also facilitate the computation of model complexity and fit statistics for non-nested models [28].

In each model, the utility weight from the EQ-5D is treated as our dependent variable. Our covariates were the summary scores from each of the 19 domains of the EORTC QLQ-C30 and QLQ-MY20, which we treated as continuous variables, in addition to respondent-level covariates such as age and gender. In total, then there are up to 22 covariate parameters in our models, being a constant term, 19 EORTC dimensions, and up to 2 demographic covariates.

The data set described in the previous section was split into a learning sample, which consisted of data from a 75% derivation sample of patients and was used to fit the different models. A validation sample was formed and consists of the remaining patients and was used to assess the predictive ability of the different candidate models.

[1] http://clincancerres.aacrjournals.org/content/early/2013/08/30/1078-0432.CCR-12-3211

Whilst our approach selects derivation/validation by patient, for simplicity we index by observations (i = 1, 2, …, 3184). Had missingness been defined at a patient level, this would have led to complete data at a point in time being removed and classified as having 'missing' data, on the basis that missing data exists at a different time point.

Four different TPMs of increasing complexity were fitted to the data:

1) TPM with second part normal regression,
2) TPM with second part normal regression with variance a function of participant characteristics,
3) TPM with second part lognormal regression, and
4) TPM with second part gamma regression with a log link.

All models considered were developed within a Bayesian framework using 'vague' prior distributions throughout and implemented in the freely available WinBUGS software [27]. In this application, Bayesian methods provided a useful tool for fitting complex, non-standard models. For a more in-depth description of Bayesian methods and their application in healthcare see [28].

3.1 Model development

As already mentioned above, previous evidence suggests that the EQ-5D score is skewed toward the upper bound. If the EQ-5D utility of observation i is Y_i then we define disutility as $D_i = 1 - Y_i$. In TPMs, we are interested in first estimating the likelihood of $D_i = 0$ (alternatively Y_i=1) and secondly how large D_i in the case it is nonzero. By doing this, TPMs essentially attempt to filter off individuals with zero-disutility, and then a distribution to the remainder (i.e. to those individuals who have a nonzero disutility).

Algebraically, let d_i be a dummy variable, which takes value 1 when $D_i > 0$ and the value 0 otherwise. Utility, as our estimated variable of interest, is then calculated based on both the likelihood of a nonzero disutility and the expected size of any non-zero

disutility. All four models have a common functional form in the first part (Part A), as follows:

$$d_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}\,[p_i] = \alpha_0 + \alpha_1 X_{1i} + \ldots + \alpha_{21} X_{22i}, \qquad (1)$$

where p_i is the probability of observing a disutility, X_{ki} indicates values for the k covariates for individual i (i.e. 19 QLQ scores plus age and gender) and the α_k parameters are estimated within the logistic model. In these cases, the vague priors used set are as follows

$$\alpha_0, \ldots, \alpha_{21} \sim N(0, 10^6).$$

To assess the importance of addressing these potential complexities, four different (Part B) models are also fitted to estimate disutility when $D_i > 0$.

Model 1: Normally distributed Part B

For those individuals who have a nonzero disutility, Part B of the model becomes:

$$(D_i \mid D_i > 0) \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{21} X_{21i} \qquad (2)$$

where X_is indicate observation-specific values for the 21 covariates (i.e. 19 QLQ scores, age and gender), and β_k 's are the regression parameters to be estimated. Vague prior distributions were specified as follows:

$$\beta_0, \ldots \beta_{21} \sim N(0, 10^6), \qquad \sigma^2 \sim \text{InverseGamma}(0.001, 0.001)$$

Model 2: Normally distributed Part B with variance a function of age and gender,

In Model 2, the TPM Model 1 again has a normal conditional distribution but allows variance to be a function of participant characteristics. In this model we allowed the logarithm of the variance to also be a function of age (X_{20i}) and gender (X_{21i}); that is:

$$(D_i \mid D_i > 0) \sim N(\mu_i, \sigma_i^2)$$

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{21} X_{21i}$$

$$\text{Log}(\sigma_i^2) = \theta_0 + \theta_1 X_{20i} + \theta_2 X_{21i} \qquad (3)$$

where again X_{ki}s indicate individual values for the 21 covariates and β's are the estimated parameters. Prior distributions were specified as:

$$\beta_0,...\beta_{21},\theta_0,\theta_1,\theta_2 \sim \mathrm{N}(0,10^6)$$

Model 3: Lognormally distributed Part B

In Model 3, the disutility is estimated with a similar equation to Model 1 but assumes a lognormal distribution with uniform variance across observations. Here, Part B is:

$$(\mathrm{D_i} \mid \mathrm{D_i} > 0) \sim \mathrm{LN}(\mu_i,\sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 \mathrm{X_{1i}} + ... + \beta_{22} \mathrm{X_{22i}} \qquad (4)$$

where X_{ki}s indicate observation-specific values for the 21 covariates (i.e. 19 QLQ scores, age and gender), and β_ks are the regression parameters to be estimated. Vague prior distributions were specified as follows:

$$\beta_0,...\beta_{21} \sim \mathrm{N}(0,10^6), \qquad \sigma^2 \sim \mathrm{InverseGamma}(0.001,0.001)$$

Model 4: Gamma distributed Part B with a log link.

The fourth model specified disutility as a Gamma regression with a log link function. Here, the log of mean disutilities is modelled as a function of the covariates. Both the mean value by observation (μ_i) and the common variance (σ^2) are related to the parameters of the Gamma distributions (a_i, b_i). (Here, $\mu_i = a_i / b_i$ and $\sigma^2 = a_i / b_i^2$)

In this model, Part B is:

$$(\mathrm{D_i} \mid \mathrm{D_i} > 0) \sim \mathrm{Gamma}\ (a_i, b_i)$$

$$a_i = \mu_i b_i \text{ and } b_i = \mu_i / \sigma^2$$

$$\log(\mu_i) = \beta_0 + \beta_1 \mathrm{X_{1i}} + ... + \beta_{21} \mathrm{X_{21i}} \qquad (5)$$

As before, the 22 parameters β_k are estimated in the log-linear model and include the constant term, 20 QLQ scores, and covariates for age and gender. Prior distributions were specified as:

$$\beta_0,...\beta_{21} \sim \mathrm{N}(0,10^6), \qquad \sigma^2 \sim \mathrm{InverseGamma}(0.001,0.001).$$

3.2 Model prediction

To assess the predictive ability of Models 1-4 each was applied to derive utilities for both the derivation and validation samples of the cohort of individuals with myeloma. The expected deficit in HRQoL is expressed as the expected mean disutility (in the case that one is observed in a particular observation) multiplied by the probability of observing a nonzero disutility, and is given by the equations (6) - (8) below:

As a similar/common first part appears across the models, we first define the expected likelihood of a disutility as:

$$p_i = P(D_i > 0) = \frac{\exp(\alpha_0 + \alpha_1 X_{1i} + ... + \alpha_k X_{ki})}{1 + \exp(\alpha_0 + \alpha_1 X_{1i} + ... + \alpha_k X_{ki})},$$

Models 1 and 2:

For the second part of the TPMs in Models 1 and 2, the conditional expectation $E(D_i \mid D_i > 0)$ from the second stage (2) equals:

$$E(D_i \mid D_i > 0) = \beta_0 + \beta_1 X_{1i} + ... + \beta_{21} X_{21i} + E(\varepsilon_i \mid D_i > 0),$$

In TPMs, this last term is assumed to equal zero in contrast to other approaches, for example, Heckman's [29] sample selection model. The relative merits of the TPM have been the subject of a vigorous debate in the literature [30] and much of the discussion focuses on this assumption. Thus,

$$E(D_i \mid D_i > 0) = \beta_0 + \beta_1 X_{1i} + ... + \beta_{21} X_{21i}.$$

Note that the unconditional expectation of the dependent variable D_i, $E(D_i)$, is given by

$$E(D_i) = P(D_i > 0).E(D_i \mid D_i > 0) + P(D_i = 0).E(D_i \mid D_i = 0).$$

Then, since $E(D_i \mid D_i = 0) = 0$, it follows that

$$E(D_i) = P(D_i > 0).E(D_i \mid D_i > 0) = (\beta_0 + \beta_1 X_{1i} + ... + \beta_{21i} X_{21i})\, p_i.$$

Therefore, for Models 1 and 2, predicted health (as one minus disutility) is:

$$\hat{Y}_i = 1 - (\beta_0 + \beta_1 X_{1i} + ... + \beta_{21i} X_{21i})\, p_i \qquad (6)$$

Model 3:

In a similar way, Model 3's lognormal formula for estimated disutility yields an expected health of:

$$\hat{Y}_i = 1 - [\exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_{21} X_{21i}) + \tfrac{1}{2}\sigma^2]\, p_i \qquad (7)$$

And for Model 4:

$$\hat{Y}_i = 1 - \exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_{21} X_{21i})\, p_i \qquad (8)$$

The performance of all models was compared by calculating the proportion of variance they explained in both the derivation and validation samples, using the unadjusted R^2 and adjusted R^2 statistics measuring the goodness of fit between the predicted and observed values [9,31].

As the intended purpose of our TPM models is to predict the mean EQ-5D scores based on the mean EORTC QLQ C-30/QLQ-MY20 domain scores, we compare the predicted versus observed mean EQ-5D scores in the overall validation data set and calculate the mean absolute prediction error. The validity of candidate models is also estimated using root mean square error (RMSE) criterion for the mean:

$$\text{RMSE} = \left(\sum_{i=1}^{T} (y_i - \hat{Y}_i)^2 / T\right)^{1/2} \qquad (9)$$

where T is the number of observations in the test sample.

Finally, the performance of the models fitted above was also compared by calculating the Bayesian Deviance Information Criterion (DIC) [28] which combines measures of both model fit and model complexity. It is defined by,

$$\text{DIC} = \bar{D} + P_D$$

where $\bar{D}$ is the posterior mean deviance and P_D is a measure of model complexity which may be termed the effective number of parameters. The DIC is similar to Akaike Information Criterion (AIC) [32] and is interpreted as a Bayesian measure of fit penalised for increased model complexity. The minimum DIC denotes the model best fitting the data [28].

4. Results

4.1. Study cohort

The entire study population consisted of 3184 observations (1839 patients); 331 observations (10%) had the EQ-5D utility weight missing and 252 observations (8%) had missing values for at least one component of EORTC QLQ-C30/QLQ-MY20. Overall, 510 observations had at least one of EQ-5D, EORTC QLQ-C30/QLQ-MY20 or demographic data values missing. The complete case analysis was therefore performed on a cohort of 2674 observations (i.e. 3184-510=2674 observations from 1658 patients); the derivation sample had 2003 observations (1244 patients) and the validation sample had 671 observations (414 patients).

Table 1 presents the baseline characteristics of the entire population, the complete case cohort, and our validation and derivation set. The mean age of the entire population was 64.75 years and 59% were males, both of which were in line with the complete case cohort and the derivation and validation data sets. The observations with missing values typically came from an older and more female sample than the observations without missing values. However, the EQ-5D scores and 19 components of the EORTC QLQ-C30/QLQ-MY20 produced similar means and interquartile ranges (IQR) across both the missing and complete sets. The correlations (≤ 0.3 and > 0.3) between the EORTC QLQ-C30/QLQ-MY20 domains and the EQ-5D scores were also explored, and the results from the final column of Table 1 showed that only dyspnoea, diarrhoea and body image had weak correlations with the EQ-5D scores.

To this end, we consider the design characteristics of both the derivation and validation samples. In Table 2 we show the mean and count for the 104 EQ-5D health states observed in the derivation sample whereas Table 3 presents the corresponding results for the 79 states observed in the validation sample. 72 of the 79 states in the validation sample also appeared in the derivation sample, with these states accounting for 99% of the validation dataset. This substantial overlap suggests that we should expect substantially similar fit between the validation and derivation samples.

4.2. Model estimation

For each model, a burn-in of 10000 iterations was allowed to reach convergence. Convergence was assessed by examining the Gelman-Rubin convergence statistic for two MCMC sequences with different initial values [33]. These were followed by a further 20,000 iterations for parameter estimation purposes.

Tables 4 and 5 summarise each the estimated distributions for the parameters (including the 19 QLQ coefficients) in terms of mean and associated 95% credible intervals. Table 4 presents the model predicting whether or not a person is likely to report full health/zero disutility, which is used as the first part of the two part models for Models 1-4[2]. Of the 19 coefficients on QLQ domains (α_1 to α_{19}), 12 have credible intervals including 0, with zero disutility associated with higher scores on emotional functioning, fatigue, physical functioning, and role functioning, but lower scores on appetite loss, disease symptoms and pain. Table 5 presents the equivalent distributions for the various Part B models. In only five domains did the coefficient's credible intervals exclude zero in all models; here, higher EQ-5D scores (lower disutility) was associated with higher scores for emotional functioning, physical functioning, future perspectives and fatigue but lower scores for disease symptoms and pain. The 95% credible intervals for age and gender included zero in all four models.

4.3. Model reliability and validation

Table 6 presents the R^2, adjusted R^2, mean prediction error statistics and RMSE for each model. No model clearly dominates across both derivation and valuation tasks. Models 1 and 2 (normal Part B with/without variance influenced by age and gender) perform best within the derivation dataset under R^2/adjusted-R^2 criteria, and explain around 70% of variation; the best performing model on the validation dataset is the TPM with Gamma regression and log link, which explains around 68% of the data. The TPM with Gamma regression and log link also provides the best performance based on i) lowest root mean squared error within the validation dataset and ii) lowest mean predicted error in the two datasets. Finally, the DIC, used to assess complexity and fit of the models to the derivation sample was also explored, and the TPM with second part

[2] Models 1-4 were estimated individually as joint Part A and Part B. However, the four different "Part A" regressions led to very similar results and each had the same parameter estimates (including estimates of standard error) up to 5 decimal places. Thus, it suffices to just state one. For completeness, Part A of Model 1 was reported in Table 4.

lognormal regression was found (Table 7) to provide the best fit to the data (DIC = -943). The TPM with Gamma regression and log link was also found to provide a good fit to the data though (DIC = -890.2).

5. Discussion

The aim of this paper was to enable the estimation of EQ-5D health state values based on the EORTC QLQ-C30 data using a Bayesian framework. We have developed a series of Bayesian TPM regression models and have found that the TPM model with normal regression and the TPM model with normal regression, with variance a function of participant characteristics perform best within the derivation dataset under R^2/adjusted R^2 criteria, and explains around 70% of variation; the best performing model on the validation dataset was the TPM with Gamma regression and log link, which explains around 68% of the data. A key advantage of the TPMs presented here is that the zero utility observations are separated from the non-zero utility observations, thus removing the need for a transformation before a standard regression model can be fitted. If zero disutility and non-zero disutility responses are believed to come from different data generating processes it is then possible to explore the determinants of these by including different sets of covariates in the two parts of the model.

All four models presented here used 'vague' prior distributions and were implemented using MCMC methods in the software package WinBUGS. This environment provides significant flexibility in model specification as it fits nonstandard models, such as those TPM models presented in this paper, in a single modelling framework, and all parameter estimation uncertainty is automatically incorporated into the results [34]. Furthermore, the DIC for model selection is also available as its computation has been coded into WinBUGS. It would have been possible to use Bayes factors instead to quantify the relative ability of the four models in predicting the data [35]. However, in comparison to DIC, the use of Bayes factors require informative prior distributions, which we did not have here.

The Tobit Model is another approach to address data with ceiling effects [36]. This approach may be considered more efficient than the TPM unless the normality and

homoscedasticity assumptions are violated [37, 38]. It is worth mentioning that Tobit model was also fitted to the EORTC QLQ-C30 data (results not shown) and residuals were examined against the predictor variables. We have found that the residuals were not constant across the level of each predictor and so our data empirically suggest that the Tobit model at least violates the assumption of homoscedasticity. Latent Class and Heckman sample selection models are also considered alternative approaches to the TPM to address data with ceiling effects. However, there is a well-established debate in health econometrics over the merits of the latter versus TPM models as of which of these works best empirically [39].

There is a scope for using more complex models such as generalized linear models [40, 41] and survival-type models to predict utilities. The latter is attractive due to the lack of assumptions required regarding the error [14] as well as their potential to cope with censored data. It is worth mentioning that implementing such models in an MCMC setting and using WinBUGS would be possible. It is perhaps worth mentioning that a TPM with second part beta regression was also fitted to the EORTC QLQ-C30 data (results not shown but available upon request from the leading author), but didn't provide further improvement over the four models presented here. We believe that the explanation for this finding is due to only 8% of the patients had EQ-5D scores at the ceiling of one. Indeed, in a healthy population, where a substantially higher proportion may have been at the ceiling, models such as TPM with second part beta regression or even models presented here may have better performance. Additional work includes dealing with missing follow-up utility data which, again, could be incorporated and implemented in MCMC setting using the Bayesian multiple imputation approach as mentioned by Kharroubi et al [42].

This paper has proposed four alternative TPM models for modelling and predicting utilities. Although it is not possible to recommend one particular model for analysing utility data in general, due to the specific characteristics of each data set and therefore the need for a series of different models to be fitted and model fit assessed, the analyses presented have demonstrated how utility data may be straightforwardly modelled using Bayesian hierarchical models, and model fit and complexity assessed using the DIC, which is straightforward to compute in a MCMC analysis. Such models provide

important information for the planning of future services and budgets, and may also be used to inform cost-effectiveness analyses.

In conclusion, we found that mean EQ-5D utility weights can be accurately estimated using a TPM regression mapping algorithm from the EORTC QLQ-C30/QLQ-MY20. Whilst previous models for mapping the EORTC QLQ-C30 to the EQ-5D exist [43, 44], this is the first model to our knowledge to explicitly consider a myeloma subgroup and to include the MY-20 data. Such a model will be of significant use to investigators conducting economic evaluations, by generating preference-based utility weights in patients with myeloma.

References

1. Brooks R. EuroQol: the current state of play. Health Policy. 1996 Jul;37(1):53-72.
2. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 2002;21(2):271-92.
3. McCabe C, Stevens K, Roberts J, Brazier JE. Health state values for the HUI2 descriptive system: results from a UK Survey. Health Economics. 2005; 14:231-244.
4. Barton GR, Sach TH, Doherty M, Avery AJ, Jenkinson C, Muir KR. An assessment of the discriminative ability of the EQ-5Dindex, SF-6D, and EQ VAS, using sociodemographic factors and clinical conditions. European Journal of Health Economics 2008;9:237–49.
5. Dan AA, Kallman JB, Srivastava R, Younoszai Z, Kim A, Younossi ZM. Impact of chronic liver disease and cirrhosis on health utilities using SF-6D and the health utility index. Liver Transplant 2008;14:321–6.
6. Wee HL, Cheung YB, Loke WC, Tan CB, Chow MH, Li SC, Fong KY, Feeny D, Machin D, Luo N, Thumboo J. The association of body mass index with health-related quality of life: an exploratory study in a multi-ethnic Asian population. Value in Health 2008;11(Suppl. 1):S105–14.
7. Kieschnick R, McCollough B. Regression analysis of variates observed on (0,1): percentages, proportions and fractions. Statistical Modelling 2003;3: 193–213.
8. Pullenayegum EM, Tarride JE, Xie F, Goeree R, Gerstein HC, O'Reilly D. Analysis of health utility data when some subjects attain the upper bound of 1: are Tobit and CLAD models appropriate? Value in Health 2010;13:487–94.
9. Huang IC, Frangakis C, Atkinson MJ, Willke R, Leite W, Vogel WB, Wu A. Addressing ceiling effects in health status measures: a comparison of techniques applied to measures for people with HIV disease. Health Service Research 2008;43:327–39.
10. Austin PC. A comparison of methods for analyzing health-related quality-of-life measures. Value in Health 2002;5:329–37.

11. Shaw JW, Pickard AS, Yu S, Iannacchione VG, Johnson JA, Coons SJ. A median model for predicting United States population-based EQ-5D health state preferences. Value in Health 2010;13:278–88.
12. Chang B-H, Pocock S. Analyzing data with clumping at zero: An example demonstration. Journal of Clinical Epidemiology 2000;53: 1036–1043.
13. Cooper NJ, Sutton AJ, Mugford M, Abrams KR. Use of Bayesian Markov Chain Monte Carlo methods to model cost-of-illness data. Medical Decision Making 2003;23: 38–53.
14. Lipscomb J, Ancukiewicz M, Parmigiani G, Hasselblad V, Samsa G, Matchar DB. Predicting the cost of illness: a comparison of alternative models applied to stroke. Medical Decision Making 1998;18: S39–S56.
15. Tooze JA, Grunwald GK, Jones KH. Analysis of repeated measures data with clumping at zero. Statistical Methods in Medical Research 2002;11: 341–355.
16. Lui, L., Strawderman R.L., Cowen M.E., Shih Y.C. A flexible two-part random effects model for correlated medical costs. Journal of Health Economics 2010;29(1): 110--23.
17. Basu A. and Manca A. Regression estimators for generic health-related quality of life and quality-adjusted life years. Medical Decision Making 2012;32(1): 56-69.
18. National Institute for Health and Clinical Excellence Guide to the methods of technology appraisal, June 2008.
19. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy, 1990. 16(3): p. 199-208.
20. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technology Assessment 1999;3:9.
21. Dolan P. Modelling valuation for EuroQol health states. Medical Care 1997;35:351–63.
22. Karlsson JA, Nilsson JÅ, Neovius M, Kristensen LE, Gülfe A, Saxne T, Geborek P. National EQ-5D tariffs and quality-adjusted life-year estimation: comparison of UK, US and Danish utilities in south Swedish rheumatoid arthritis patients. Annals of the Rheumatic Disease 2011;70(12):2163-6.
23. Bottomley A, Aaronson NK. International perspective on health related quality-of-life research in cancer clinical trials: the European Organisation for Research

and Treatment of Cancer experience. Journal of Clinical Oncology 2007;25: 5082–6.

24. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC. The European Organization for Research and Treatment of Cancer QLQ -C30: a quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 1993; 85:365–76.
25. Stead, M., Brown, J., Velikova, G., Kaasa, S., WisslØff, F., Child, J., Hippe, E., Hjorth, M., Sezer, O., Selby, P. Development of an EORTC questionnaire module to be used in health-related quality-of-life assessment for patients with multiple myeloma. British Journal of Haematology, 1999, 104, 605–611.
26. Morgan GJ, Davies FE, Gregory WM, Cocks K, Bell SE, Szubert AJ, et al. First-line treatment with zoledronic acid as compared with clodronic acid in multiple myeloma (MRC Myeloma IX): a randomised controlled trial. Lancet 2010;376:1989–99.
27. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual: Version 1.4. MRC Biostatistics Unit: Cambridge 2003.
28. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 2002;64(4): 583–639.
29. Heckman, J. Sample selection bias as a specification error Econometrica, 47;1979, pp. 53–161.
30. Dow WH. and Norton EC. Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions. Health Services & Outcomes Research Methodology 2003;4, 5-18.
31. Lawrence WF, Fleishman JA. Predicting EuroQoL EQ-5D preference scores from the SF-12 Health Survey in a nationally representative sample. Medical Decision Making. 2004;24(2):160–9.
32. Akaike H. Information theory and an extension of the maximum likelihood principle. In Procedures 2nd International Symposium Information Theory, Petrov BN, Caski F (eds). Akademiai Kiado: Budapest, 1973;267–281.
33. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science 1992;7: 457–472.

34. Cooper, N.J., Lambert, P.C., Abrams, K.R., and Sutton, A.J. Predicting costs over time using Bayesian Markov Chain Monte Carlo Methods: An application to early inflammatory polyarthritis. Health Economics 2007;16, 37–56.
35. Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association 1995;90: 773–795.
36. Austin, P. C. M. Escobar, and J. A. Kopec. The Use of the Tobit Model for Analyzing Measures of Health Status. Quality of Life Research 2000;9 (8): 901–10.
37. Kennedy, P. A Guide to Econometrics. Cambridge, MA: MIT Press 1998.
38. Greene, W.H. Econometric Analysis. Upper Saddle River, NJ: Prentice Hall 2003.
39. Madden, D. Sample selection versus two-part models revisited: The case of female smoking and drinking. Journal of Health Economics 2008;27(2): 300-307.
40. Pullenayegum EM, Wong HS, Childs A. Generalized additive for the analysis of EQ-5D utility data. Medical Decision Making 2013; 33(2): 244-51.
41. Hernández Alava M, Wailoo AJ, Ara R. Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. Value Health. 2012 May;15(3):550-61.
42. Kharroubi, SA, Meads D, Edlin R, Browne C, McCabe C. Use of Bayesian Markov Chain Monte Carlo methods to estimate EQ-5D utility scores from EORTC QLQ data in Myeloma. Medical Decision Making 2015; 35(3), 351-360.
43. Doble B, Lorgelly P. Mapping the EORTC QLQ-C30 onto the EQ-5D-3L: assessing the external validity of existing mapping algorithms. Quality of Life Research. 2016; 5(4):891-911.
44. Khan I, Morris S. A non-linear beta-binomial regression model for mapping EORTC QLQ- C30 to the EQ-5D-3L in lung cancer patients: a comparison with existing approaches. Health Quality Life Outcomes. 2014; 12;12-163.

Table 1: Baseline Characteristics of Study Cohort (Kharroubi et al [45]).

		Missing Data		Complete			
	Entire population	EQ-5D	QLQ-MY20	Case Cohort	Derivation Set	Validation Set	ρ
No of patients	1839	318	240	1658	1244	414	
No of obn	3184	331	252	2674	2003	671	
Age, mean(IQR)	64.75 (58-72)	65.6 (58-73)	67.53 (60-76)	64.45 (58-72)	64.4 (58-72)	64.62 (58-72)	
Male, %	59.22	56.6	55	59.89	59.97	59.66	
EQ-5D, mean(IQR)	0.52(0.26-0.76)	NA	0.47(0.19-0.76)	0.52(0.26-0.76)	0.52(0.26-0.76)	0.52(0.26-0.76)	
PF, mean(IQR)	58.05(40-80)	58.11(40-80)	NA	58.29(40-80)	58.02(40-80)	58.08(40-80)	0.75
RF, mean(IQR)	42.16(0-66.67)	41.53(0-66.67)	NA	42.40(0-66.67)	42.17(0-66.67)	42.09(0-66.67)	0.68
DY, mean(IQR)	32.24(0-66.67)	32.43(0-66.67)	NA	32.01(0-33.34)	32.10(0-33.34)	31.74(0-66.67)	-0.26
PA, mean(IQR)	44.01(16.67-66.67)	48.26(16.67-83.34)	NA	43.46(16.67-66.67)	43.33(16.67-66.67)	43.84(16.67-66.67)	-0.70
FA, mean(IQR)	51.52(33.34-66.67)	51.20(33.34-66.67)	NA	51.45(33.34-66.67)	51.34(33.34-66.67)	51.79(33.34-66.67)	-0.61
SL, mean(IQR)	34.77(0-66.67)	35.47(0-66.67)	NA	34.60(0-66.67)	34.47(0-66.67)	35.02(0-66.67)	-0.35
AP, mean(IQR)	29.68(0-66.67)	30.73(0-66.67)	NA	29.24(0-66.67)	29.51(0-66.67)	28.47(0-33.34)	-0.45
NV, mean(IQR)	13.20(0-16.67)	14.81(0-16.67)	NA	12.93(0-16.67)	13.13(0-16.67)	12.34(0-16.67)	-0.36
CO, mean(IQR)	32.41(0-66.67)	35.84(0-66.67)	NA	32.11(0-66.67)	31.84(0-66.67)	32.94(0-66.67)	-0.39
DI, mean(IQR)	10.07(0-0)	11.95(0-0)	NA	9.81(0-0)	10.18(0-0)	8.70(0-0)	-0.10
CF, mean(IQR)	73.10(50-100)	72.53(50-100)	NA	73.22(66.67-100)	73.13(66.67-100)	73.50(66.67-100)	0.48
EF, mean(IQR)	70.24(58.34-91.67)	70.09(55.56-91.67)	NA	70.38(58.34-91.67)	70.51(58.34-91.67)	70.00(58.34-91.67)	0.51
SF, mean(IQR)	50.11(16.67-83.34)	50.86(16.67-83.34)	NA	50.09(16.67-83.34)	50.08(16.67-83.34)	50.10(16.67-83.34)	0.64
FI, mean(IQR)	20.54(0-33.34)	19.51(0-33.34)	NA	20.61(0-33.34)	21.32(0-33.34)	18.48(0-33.34)	-0.30
QL, mean(IQR)	51.15(33.34-66.67)	48.93(33.34-66.67)	NA	51.53(33.34-66.67)	51.46(33.34-66.67)	51.71(33.34-66.67)	0.64
DS, mean(IQR)	31.91(13.34-50)	34.18(16.67-50)	NA	31.59(11.12-46.67)	32.04(11.12-50)	30.22(13.34-44.45)	-0.62
SE, mean(IQR)	23.77(11.12-33.34)	24.15(11.12-33.34)	NA	23.66(11.12-33.34)	23.57(11.12-33.34)	23.93(11.12-33.34)	-0.51
BI, mean(IQR)	70.98(33.34-100)	73.15(66.67-100)	NA	71.09(33.34-100)	71.72(33.34-100)	67.20(33.34-100)	0.28
FP, mean(IQR)	48.16(33.34-66.67)	46.71(22.23-66.67)	NA	48.33(33.34-66.67)	48.95(33.34-66.67)	45.49(22.23-66.67)	0.35

Note: PF, physical functioning; RF, role functioning; DY, dyspnoea; PA, pain; FA, fatigue; SL, insomnia; AP, appetite loss; NV, Nausea and vomiting; CO, constipation; DI, diarrhoea; CF, cognitive functioning; EF, emotional functioning; SF, social functioning; FI, financial difficulties; QL, quality of life; DS, disease symptoms; SE, side effects; BI, body image; FP, future perspective; NA, not applicable; ρ, correlation coefficient.

Table 2: Mean and count for 104 health states in the derivation sample

State	Mean	n	State	Mean	n	State	Mean	n	State	Mean	n
11111	1	169	12232	0.053	1	21332	0.03	7	23322	0.079	8
11112	0.848	62	12311	0.452	1	21333	-0.135	2	23323	-0.086	1
11113	0.414	3	12321	0.329	3	22111	0.746	1	23332	-0.184	8
11121	0.796	88	12322	0.258	2	22132	0.02	1	23333	-0.349	4
11122	0.725	57	12331	0.066	1	22211	0.71	9	31311	0.242	2
11123	0.291	1	12332	-0.005	3	22212	0.639	9	31312	0.171	1
11211	0.883	55	21111	0.85	21	22213	0.205	1	31322	0.048	1
11212	0.812	21	21112	0.779	10	22221	0.587	64	31323	-0.117	1
11221	0.76	115	21121	0.727	34	22222	0.516	117	32131	-0.154	1
11222	0.689	68	21122	0.656	12	22223	0.082	10	32222	0.002	1
11223	0.255	8	21131	0.195	3	22231	0.055	20	32311	0.138	5
11232	0.157	4	21211	0.814	38	22232	-0.016	39	32312	0.067	2
11311	0.556	4	21212	0.743	22	22233	-0.181	7	32313	-0.098	1
11312	0.485	2	21221	0.691	164	22311	0.383	9	32321	0.015	1
11321	0.433	14	21222	0.62	145	22312	0.312	9	32322	-0.056	10
11322	0.362	11	21223	0.186	8	22313	0.147	3	32323	-0.221	1
11323	0.197	1	21231	0.159	18	22321	0.26	48	32331	-0.248	1
11331	0.17	1	21232	0.088	28	22322	0.189	119	32332	-0.319	5
11332	0.099	2	21233	-0.077	1	22323	0.024	7	32333	-0.484	3
11333	-0.066	2	21311	0.487	12	22331	-0.003	22	33311	0.028	2
12112	0.744	1	21312	0.416	8	22332	-0.074	61	33312	-0.043	1
12132	0.089	1	21313	0.251	2	22333	-0.239	9	33321	-0.095	2
12213	0.274	1	21321	0.364	31	23222	0.137	1	33322	-0.166	4
12221	0.656	7	21322	0.293	45	23231	-0.055	1	33331	-0.358	3
12222	0.585	10	21323	0.128	5	23311	0.273	1	33332	-0.429	14
12231	0.124	2	21331	0.101	6	23321	0.15	2	33333	-0.594	7

Table 3: Mean and count for 79 health states in the validation sample

State	Mean	n	State	Mean	N	State	Mean	n
11111	1	51	21132	0.124	1	22313	0.147	1
11112	0.848	21	21211	0.814	16	22321	0.26	13
11121	0.796	30	21212	0.743	8	22322	0.189	38
11122	0.725	18	21221	0.691	51	22323	0.024	5
11131	0.264	2	21222	0.62	54	22331	-0.003	4
11211	0.883	20	21223	0.186	3	22332	-0.074	19
11212	0.812	16	21231	0.159	8	22333	-0.239	2
11213	0.378	1	21232	0.088	5	23222	0.137	3
11221	0.76	28	21233	-0.077	1	23311	0.273	1
11222	0.689	24	21311	0.487	4	23321	0.15	1
11223	0.255	4	21312	0.416	3	23322	0.079	6
11311	0.556	2	21321	0.364	9	23323	-0.086	1
11312	0.485	1	21322	0.293	13	23332	-0.184	3
11321	0.433	5	21323	0.128	2	31322	0.048	1
11322	0.362	3	21331	0.101	3	32313	-0.098	1
12212	0.708	2	21332	0.03	5	32321	0.015	1
12221	0.656	4	22121	0.623	1	32322	-0.056	2
12222	0.585	2	22211	0.71	3	32331	-0.248	2
12312	0.381	1	22212	0.639	3	32332	-0.319	4
12321	0.329	1	22221	0.587	32	32333	-0.484	3
12322	0.258	1	22222	0.516	50	33311	0.028	1
12332	-0.005	1	22223	0.082	1	33321	-0.095	1
13322	0.148	1	22231	0.055	4	33322	-0.166	2
21111	0.85	2	22232	-0.016	9	33332	-0.429	5
21112	0.779	1	22233	-0.181	2	33333	-0.594	1
21121	0.727	7	22311	0.383	1			
21122	0.656	8	22312	0.312	1			

Table 4: Results of first part of Models 1-- 4

	QLQ domain*	Coefficient (credible interval)
α_0 (Constant)		**15.65 (11.54, 20.10)**
α_1	**Appetite Loss**	**2.233 (0.766, 3.820)**
α_2	Body Image	-0.172 (-1.205, 0.881)
α_3	Cognitive Functioning	1.124 (-0.729, 2.810)
α_4	Constipation	-0.292 (-1.268, 0.701)
α_5	Diarrhoea	1.253 (-0.357, 2.958)
α_6	**Disease Symptoms**	**4.934 (2.235, 7.739)**
α_7	Dyspnoea	-0.332 (-1.485, 0.798)
α_8	**Emotional Functioning**	**-6.648 (-9.105, -4.415)**
α_9	**Fatigue**	**-2.105 (-4.017, -0.209)**
α_{10}	Financial Difficulties	1.145 (-0.204, 2.594)
α_{11}	Future Perspective	-0.664 (-1.818, 0.463)
α_{12}	Nausea and Vomitingg	-0.013 (-2.846, 2.908)
α_{13}	**Pain**	**5.874 (3.645, 8.201)**
α_{14}	**Physical Functioning**	**-6.703 (-9.398, -4.095)**
α_{15}	Quality of Life	-1.706 (-3.563, 0.138)
α_{16}	**Role Functioning**	**-2.929 (-4.529, -1.251)**
α_{17}	Side Effects	-0.544 (-3.521, 2.503)
α_{18}	Social Functioning	0.063 (-1.477, 1.592)
α_{19}	Insomnia/Sleep	-0.628 (-1.708, 0.471)
α_{20} (age)		-0.009 (-0.034, 0.0142)
α_{21} (female)		-0.202 (-0.740, 0.316)

Note: (*) All variables are included as 100x standard QLQ domains. Values given as posterior mean (central 95% credible interval). Estimates shown in bold are those who have credible intervals excluding zero.

Table 5: Results of second part models (Models 1-- 4)

	QLQ*	Model 1	Model 2	Model 3	Model 4
β_0 (Constant)		**0.885 (0.793, 0.972)**	**0.876 (0.782, 0.961)**	**-0.221 (-0.392, -0.047)**	**-0.223 (-0.374, -0.050)**
β_1	AP	**0.051 (0.017, 0.086)**	**0.053 (0.018, 0.086)**	0.066 (-0.001, 0.134)	**0.076 (0.013, 0.144)**
β_2	BI	0.023 (-0.005, 0.052)	0.024 (-0.005, 0.053)	0.048 (-0.008, 0.104)	0.040 (-0.023, 0.094)
β_3	CF	0.017 (-0.031, 0.064)	0.014 (-0.033, 0.060)	0.078 (-0.016, 0.170)	**0.106 (0.031, 0.182)**
β_4	CO	0.015 (-0.014, 0.044)	0.016 (-0.013, 0.045)	0.000 (-0.057, 0.057)	0.006 (-0.057, 0.054)
β_5	DI	**-0.054 (-0.097, -0.011)**	**-0.058 (-0.101, -0.016)**	-0.077 (-0.161, 0.006)	**-0.091 (-0.176, -0.011)**
β_6	DS	**0.124 (0.062, 0.185)**	**0.129 (0.067, 0.190)**	**0.208 (0.090, 0.328)**	**0.189 (0.072, 0.303)**
β_7	DY	**-0.038 (-0.071, -0.006)**	**-0.037 (-0.069, -0.005)**	-0.045 (-0.107, 0.018)	-0.050 (-0.106, 0.014)
β_8	EF	**-0.145 (-0.197, -0.092)**	**-0.144 (-0.199, -0.091)**	**-0.241 (-0.346, -0.137)**	**-0.266 (-0.378, -0.168)**
β_9	FA	**-0.088 (-0.147, -0.027)**	**-0.085 (-0.144, -0.025)**	**-0.144 (-0.261, -0.026)**	**-0.121 (-0.231, -0.011)**
β_{10}	FI	0.002 (-0.029, 0.033)	0.002 (-0.028, 0.031)	0.053 (-0.007, 0.113)	0.038 (-0.020, 0.095)
β_{11}	FP	**-0.073 (-0.115, -0.031)**	**-0.071 (-0.112, -0.029)**	**-0.144 (-0.226, -0.059)**	**-0.136 (-0.223, -0.070)**
β_{12}	NV	0.020 (-0.032, 0.071)	0.020 (-0.031, 0.072)	0.041 (-0.061, 0.143)	0.042 (-0.056, 0.140)
β_{13}	PA	**0.218 (0.171, 0.265)**	**0.217 (0.171, 0.264)**	**0.463 (0.370, 0.553)**	**0.463 (0.375, 0.557)**
β_{14}	PF	**-0.510 (-0.568, -0.452)**	**-0.515 (-0.571, -0.457)**	**-0.955 (-1.066, -0.843)**	**-0.820 (-0.909, -0.718)**
β_{15}	QF	**-0.090 (-0.153, -0.025)**	**-0.081 (-0.142, -0.020)**	-0.117 (-0.237, 0.001)	-0.119 (-0.236, 0.013)
β_{16}	RF	-0.038 (-0.086, 0.011)	-0.034 (-0.083, 0.014)	**-0.200 (-0.294, -0.105)**	**-0.257 (-0.337, -0.173)**
β_{17}	SE	**0.086 (0.002, 0.172)**	**0.092 (0.011, 0.175)**	0.113 (-0.053, 0.271)	0.141 (-0.033, 0.313)
β_{18}	SF	**-0.055 (-0.100, -0.012)**	**-0.056 (-0.101, -0.011)**	**-0.129 (-0.214, -0.044)**	**-0.139 (-0.214, -0.069)**
β_{19}	SL	0.009 (-0.022, 0.040)	0.008 (-0.022, 0.039)	0.013 (-0.046, 0.074)	0.012 (-0.046, 0.069)
β_{20} (age)		0.000 (-0.001, 0.001)	0.000 (-0.001, 0.001)	0.000 (-0.002, 0.002)	0.001 (-0.001, 0.003)
β_{21} (female)		-0.015 (-0.034, 0.04)	-0.014 (-0.033, 0.005)	-0.015 (-0.053, 0.021)	-0.007 (-0.042, 0.030)
σ		**0.191 (0.184, 0.197)**	NA	**0.372 (0.360, 0.384)**	**0.189 (0.184, 0.197)**
θ (Constant)		NA	**3.438 (3.241, 3.633)**	NA	NA
θ_1 (age)		NA	**-0.009 (-0.016, -0.003)**	NA	NA
θ_2 (female)		NA	-0.084 (-0.217, 0.047)	NA	NA

Note: PF, physical functioning; RF, role functioning; DY, dyspnoea; PA, pain; FA, fatigue; SL, insomnia; AP, appetite loss; NV, Nausea and vomiting; CO, constipation; DI, diarrhoea; CF, cognitive functioning; EF, emotional functioning; SF, social functioning; FI, financial difficulties; QL, quality of life; DS, disease symptoms; SE, side effects; BI, body image; FP, future perspective; NA, not applicable; (*) All variables are included as 100x standard QLQ domains. Values given as posterior mean (central 95% credible interval). Estimates shown in bold are those who have credible intervals excluding zero.

Table 6: Model performance based on central estimate

	Derivation data set			Validation data set			
	R^2	Adjusted R^2	Mean predicted error	R^2	Adjusted R^2	Mean predicted error	RMSE
Model 1	**0.7005**	**0.6940**	0.1431	0.6787	0.6572	0.1471	0.1892
Model 2	0.7005	0.6937	0.1430	0.6778	0.6552	0.1472	0.1894
Model 3	0.6864	0.6797	0.1463	0.6714	0.6494	0.1466	0.1913
Model 4	0.6959	0.6893	**0.1387**	**0.6803**	**0.6589**	**0.1394**	**0.1887**

Note: R^2, proportion of variance explained by the model; Estimates shown in bold are best performing models.

Table 7: Overall DIC for the fitted models

	Model 1	Model 2	Model 3	Model 4
$\bar{D}$	-337.6	-345.0	-988.2	-934.6
P_D	44.76	46.53	45.20	44.46
DIC	-292.8	-298.5	-943.0	-890.2